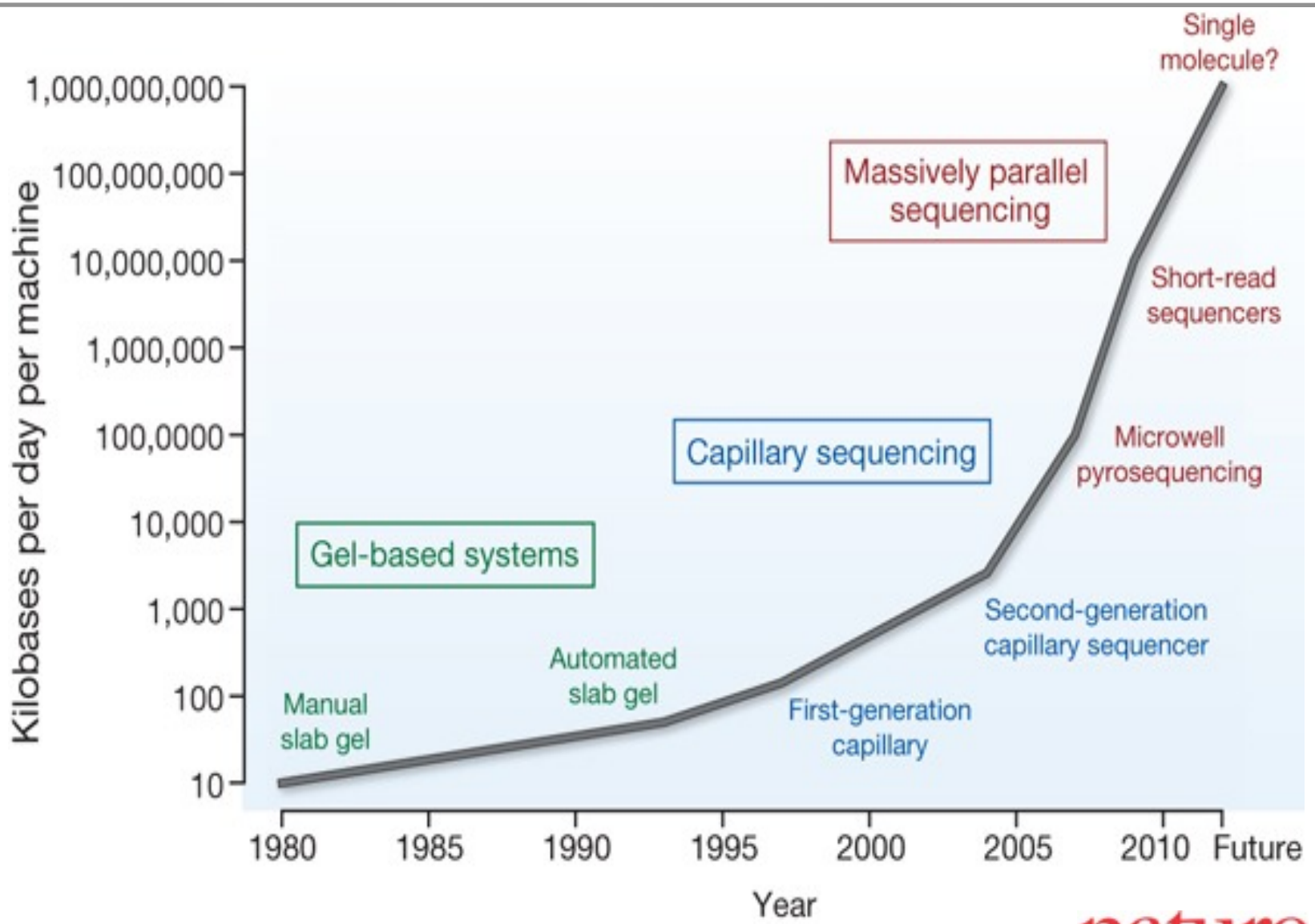
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Improvements in the rate of DNA sequencing over the past 30 years and into the future



MR Stratton *et al. Nature* **458**, 719-724 (2009) doi:10.1038/nature07943

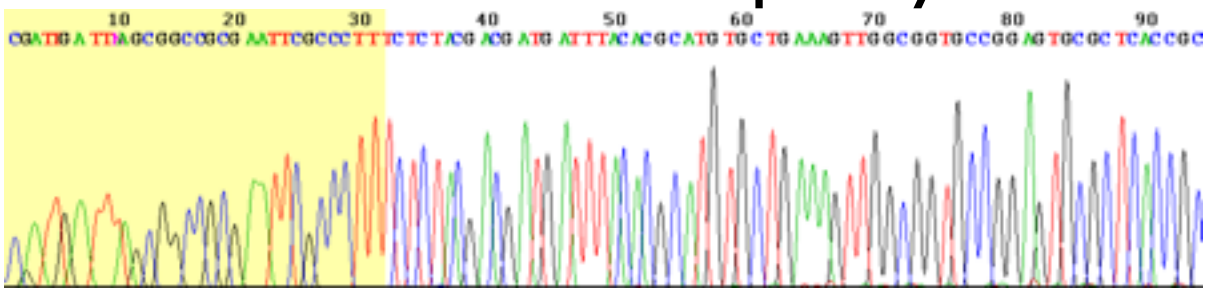
nature

First generation sequencing: Sanger chain termination

Sanger chain termination

↓

automated 4-channel capillary machines



↓

High-throughput sequencing

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

Sequencing technologies

- First generation
  - Sanger: >800 nt, low output
- Second ("next") generation
  - Roche/454 FLX: 200-500 nt, 1Gb/day
  - Illumina/Solexa: 100 nt, 25 Gb/day
  - ABI SOLiD
- Third ("next-next") generation
  - Single molecule sequencing

DTU

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

NG Sequencing

- Illumina ~6 billion reads (100 bp)

HiSeq 2000 Instrument

The HiSeq 2000 sequencing system offers unprecedented output and a breakthrough user experience. Leveraging Illumina's proven and widely-adopted, reversible terminator-based sequencing by synthesis chemistry in combination with innovative engineering, HiSeq 2000 delivers the industry's highest sequencing output and fastest data generation rate. Human interaction design features and the easiest sequencing workflow set a new standard for simplicity and user experience.

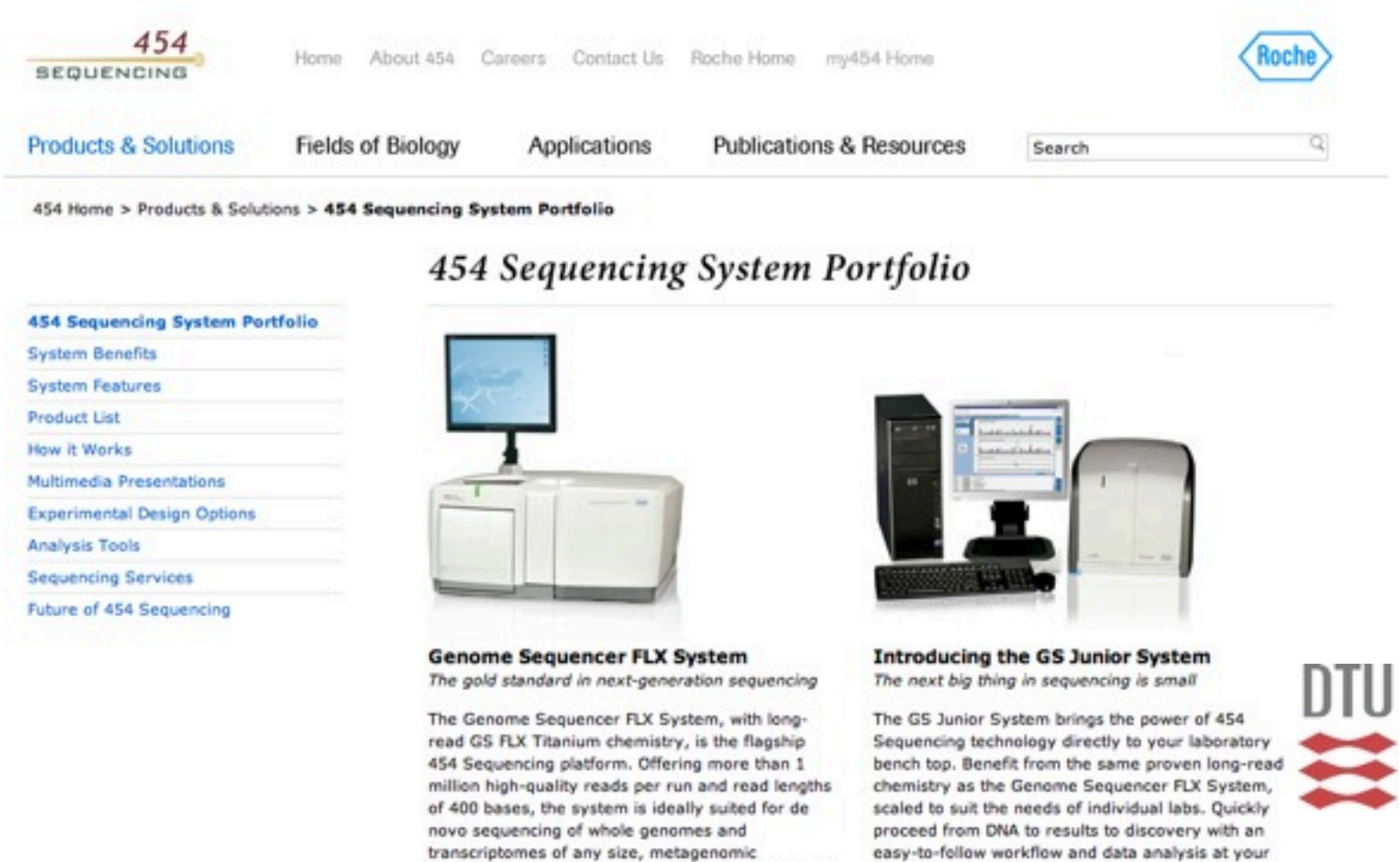
illumina®

DTU

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

NG Sequencing

- Illumina ~6 billion reads (100 bp)
- Roche 454 >1 million reads (800bp)



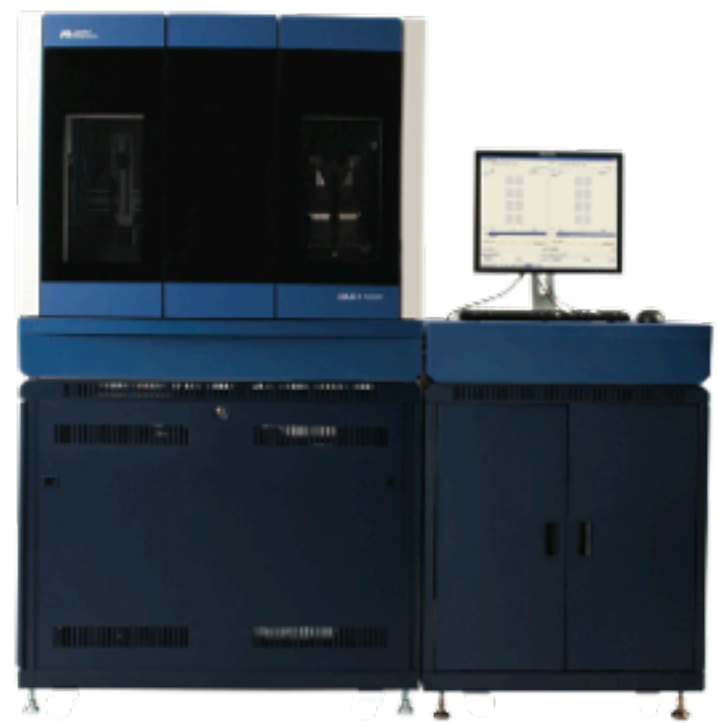
DTU

NG Sequencing

- Illumina ~6 billion reads (100 bp)
- Roche 454 >1 million reads (800bp)
- ABI Solid ~100 Gbp per run (50bp)

Key Benefits

- **Higher accuracy**—detection of causative variation enabled at lower coverage and cost per sample
- **Scalable throughput on a single platform**—80–100 GB of mappable sequence per run
- **Automated workflow**—80% reduction in hands-on time and increased reproducibility in yield allow for significant time and labor savings
- **True paired-end sequencing**—bidirectional sequencing facilitates detection of genetic alterations as well as splice variants and fusion transcripts with lower sample input
- **Robust multiplexing kits**—intelligent barcode strategy enables accurate assignment without introduction of bias



Comparison of NG sequencing technologies

	ABI SOLiD	Illumina GA	Roche 454 FLX
Cost	SOLiD 4: $495k SOLiD PI: $240k	IIe: $470k IIx: $250k HiSeq: $690k	Titanium: $500k
Quantity of Data per run	SOLiD 4: 100Gb SOLiD PI: 50Gb	IIe: 20 - 38 Gb IIx: 50 – 95 Gb HiSeq: 200Gb +	450 Mb
Run Time	7 Days	4 Days	9 Hours
Pros	*Low error rate* due to dibase probes	Most widely used NGS platform. Requires least DNA	Short run time. *Long reads* better for de novo sequencing
Cons	Long run times. Has been demonstrated certain reads don't match reference	Least multiplexing capability of the 3. Poor coverage of AT rich regions	Expensive reagent cost. Difficulty reading homopolymer regions

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

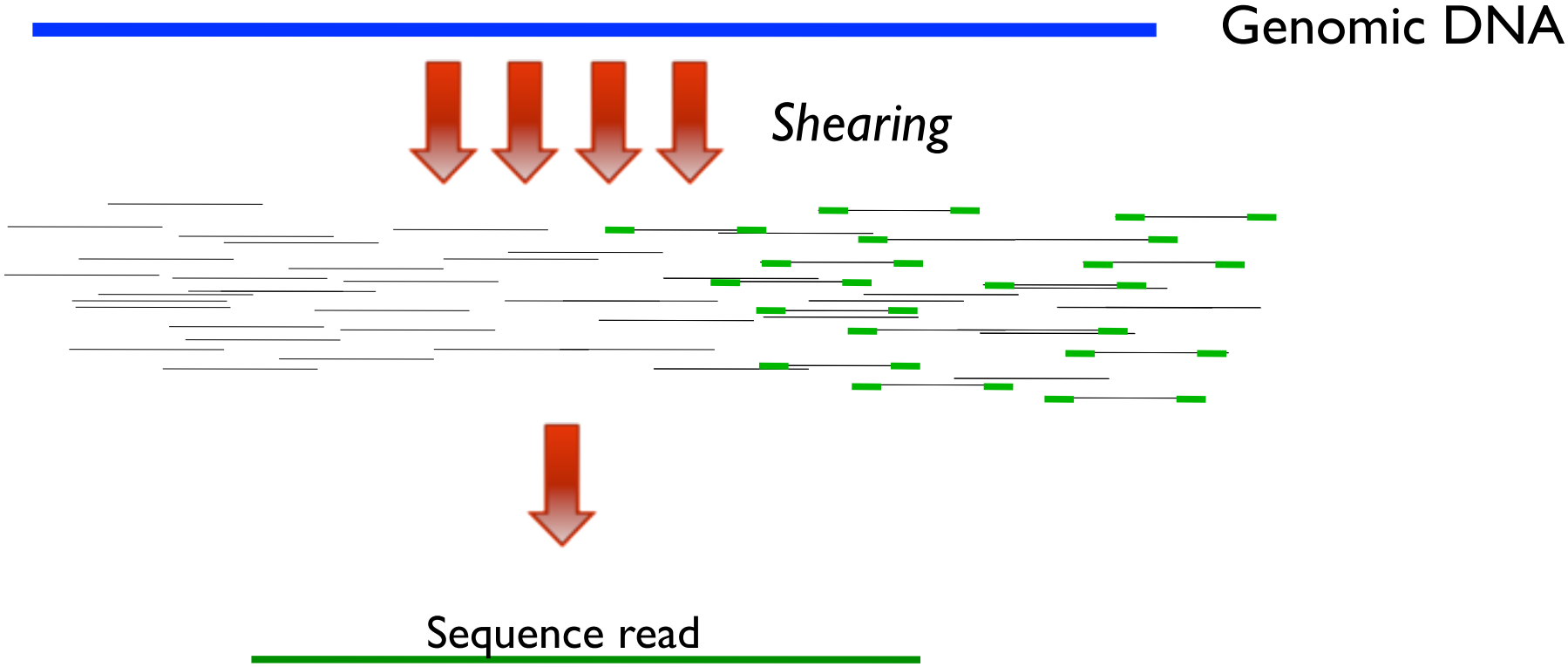
Newer sequencing technologies

- Ion Torrent (second gen.)
  - 100-200bp

- Pacific BioSciences ("third gen.")
  - ~1500bp
  - Single molecule real time
  - dye on phosphate
  - http://www.youtube.com/watch?v=_B_cUZ8hSYU

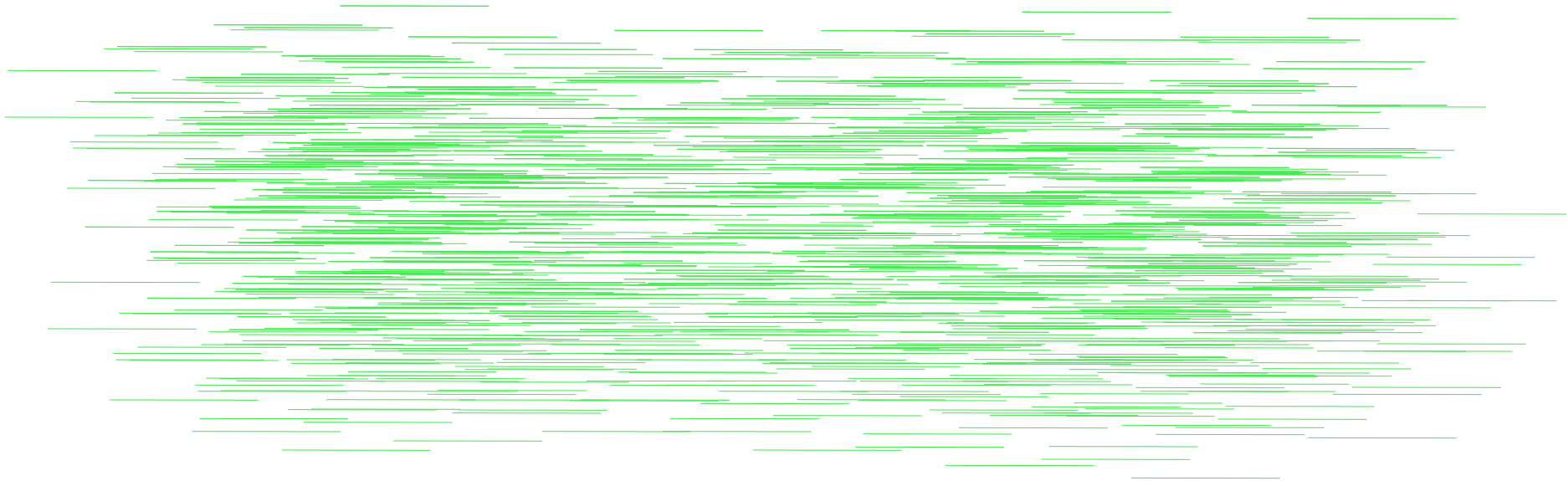
DTU

Sequencing



35-800bp

Millions of sequence reads ...



put reads together

reconstruct the complete genome

CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Sequencing data output

FASTQ format:

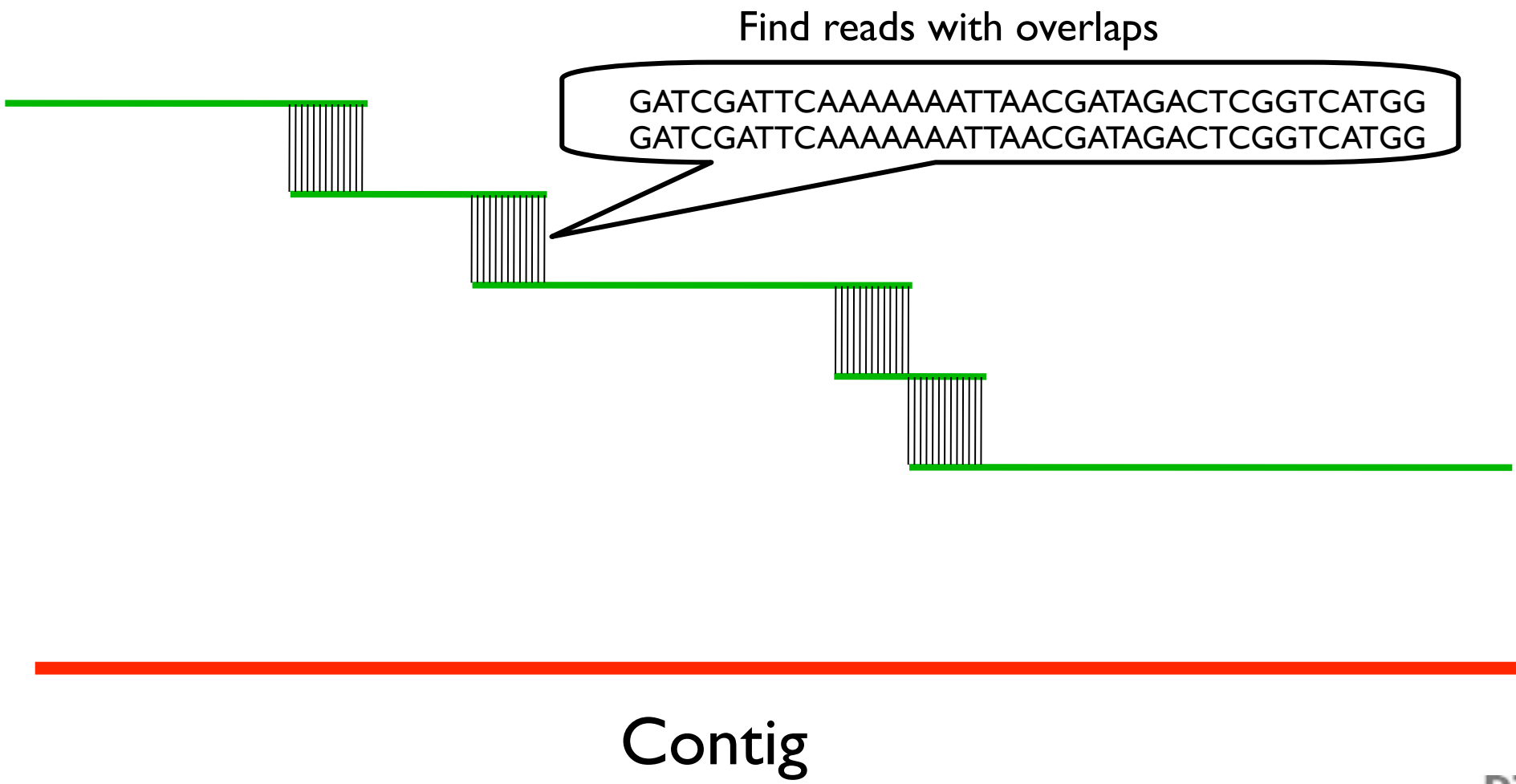
```
@MJANASCHII_25:7:1:3:1210/1
AAGATACTGCCAATATCTTGCTTGATAGTGGAATGAAAAAAGATTATTCGAAAAAACAGTCCGA
+
aaaaaaababaababababaaaaabb`]``_MZab__Y_bbbb_ZXbaSaa[__bbbb]H]YV_^_
@MJANASCHII_25:7:1:4:1887/1
TTGATAGTGGAATGAAAAAAGATTATTCGAAAAAACAGTCCGATTTGATCACAACGGCAGGATT
+
abaaa`bbaa_aab`b]``[P[abYa`^aTT_\^[YU]PWZ]Xa\MUTTNTZEMRBBBBBBBBB
@MJANASCHII_25:7:1:4:1508/1
AAAATCAAAAATTTTGGGCCCGTGTAACCGAACTGGGTGGGTGGGGTATCGGTAAACGGGGGTT
+
aabaabbbbb`baaaaa_RN\_SF]a_I[N]_BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@MJANASCHII_25:7:1:4:473/1
GTGGCGCTAGTTTAGGTAAAACGCATTTGATGCATGCAATTGGGAATCAGATTTTAACTGATAA
+
abababffgghhhhhghhfhhfa^aaba`aaaaaaa_aaa_XTZ`\U[aa[RETHXXP_aM\OT
@MJANASCHII_25:7:1:4:023/1
CCGCTGCCCGCCGGTAAGAGGCCTCCCGAGGCGCCCGCCGAAGACCGCTCCCTCGGCCGCCGCC
+
abbbbbbaababbaabbaa_`abaabaababababaaba`b``a^ababaa`ba^^a`P^^aaW
```

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

How?

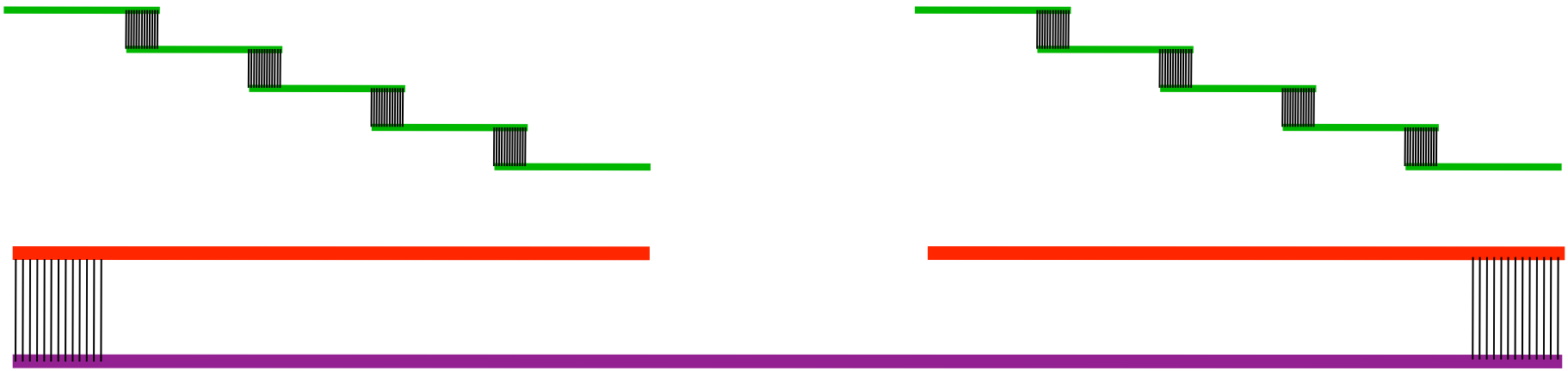
DTU

de-novo assembly



de-novo assembly

reads (paired ends, known distance)



supercontig with known direction

TATGCAGTCGATGTCCCAGATCCCCGGACAGACGGCGGGCGGAAAGATCAAACGCTTCCTTTTGCGCCGGATGGTTGGTCAGAAGATAGGTTCTGTCGTCTGCGCCCGCTTCCAGAAAATCCGAATATACCGTGTCCAGATCATGGCGGAACGCCATCAGATCTTCC

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Try it out .... de-novo assembly

Assemble the pieces together by using areas of similarity

- a minimum of a 7-bp overlap
- overlap must not include any N bases.
- same orientation so that the sequence can be read from left to right.
- There might be a single base pair difference. This base pair difference can be due to such factors as genetic polymorphism or low quality scores.
- simplified – no double-stranded DNA

Valid Assemblies

```
..NNNNGGACTATGATTCG
            |||||||
           TGATTCGAGGCTAANN..
```

```
..NNNNNNNNCGATTCTGATCCGA
          |||||||
     GTCCTCGATTCTNNNNNNNN..
```

Invalid Assemblies

```
..NNNNCGGACTATGATT
            ||||||
           ATGATTCGAGGCTAANN..
```

no 7bp overlap

```
..NNNNNNNNCGCTACTGATCCGA
          ||  |  |||
     GTCCTCGATTCTGNNNNNNNN..
```

too many mismatches

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

What if ...

We have a sequenced reference genome?

e.g. *Methanococcus janaschii*

DTU

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

Different Assembly types

- De-Novo Assembly
- Reference Genome Guided Assembly

DTU

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

Try it out .... reference-guided assembly

Map the following sequences to the reference genome

```
@MJANASCHII_25:7:1:3:1210/1
AAGATACTGCCAATATCTTGCTTGATAGTGGAATGAAAAAAGATTATTCGAAAAAACAGTCCGA
+
aaaaaaababaababababaaaaabb`]``_MZab__Y_bbbb_ZXbaSaa[__bbbb]H]YV_^_
@MJANASCHII_25:7:1:4:1887/1
TTGATAGTGGAATGAAAAAAGATTATTCGAAAAAACAGTCCGATTTGATCACAACGGCAGGATT
+
abaaa`bbaa_aab`b]``[P[abYa`^aTT_\^[YU]PWZ]Xa\MUTTNTZEMRBBBBBBBBB
@MJANASCHII_25:7:1:4:1508/1
AAAATCAAAAATTTTGGGCCCGTGTAACCGAACTGGGTGGGTGGGGTATCGGTAAACGGGGGTT
+
aabaabbbbb`baaaaa_RN\_SF]a_I[N]_BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@MJANASCHII_25:7:1:4:473/1
GTGGCGCTAGTTTAGGTAAAACGCATTTGATGCATGCAATTGGGAATCAGATTTTAACTGATAA
+
abababffgghhhhhghhfhhfa^aaba`aaaaaaa_aaa_XTZ`\U[aa[RETHXXP_aM\OT
@MJANASCHII_25:7:1:4:023/1
CCGCTGCCCGCCGGTAAGAGGCCTCCCGAGGCGCCCGCCGAAGACCGCTCCCTCGGCCGCCGCC
+
abbbbbbaababbaabbaa_`abaabaababababaaba`b``a^ababaa`ba^`a`P^`aaW
```

DTU

Tools

- De-novo assembly
    - Velvet (454, Illumina, SOLiD)
    - SOAPdenovo (Illumina)
    - Newbler (454)

- Reference mapping
    - BWA